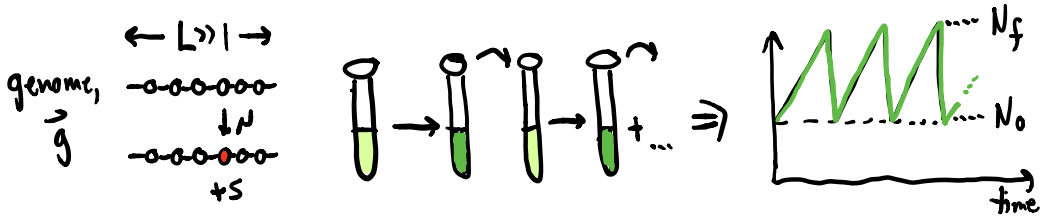
Announcements: Problem Set 4 Posted (Last one!) DUE: 3/9/21

Last time: Intro to multi-locus models of evolution



$$\frac{\partial f(\vec{g},t)}{\partial t} = \underbrace{\left(x(\vec{g}) - \bar{x}\right) f(\vec{g})}_{\text{natural selection}} + \underbrace{\sum_{\vec{g}'} M(\vec{g}' \to \vec{g}) f(\vec{g}') - M(\vec{g} \to \vec{g}') f(\vec{g})}_{\text{mutations}}$$

fitness of genotype; mean fitness of pop'n $= \bar{x}(t) = \int x(\vec{g}) f(\vec{g}) d\vec{g}$; incoming; outgoing

$$+ \sqrt{\frac{f(\vec{g})}{N}}\,\eta(\vec{g}) - f(\vec{g}) \sum_{\vec{g}'} \sqrt{\frac{f(\vec{g}')}{N}}\,\eta(\vec{g}') \Bigg\} \text{ genetic drift}$$

$\Updownarrow$

$$\frac{\partial f}{\partial t} = s f(1-f) + \mu(1-f) - \nu f + \sqrt{\frac{f(1-f)}{N}}\,\eta(t)$$

so far, similar to $L=1$ case but with more dimensions...

Today: ① two new biological features that enter for $L \geq 2$.

② How can we start to understand these models?

(4) "Epistasis": properties of $\vec{g} \to X(\vec{g})$ map
("fitness landscape")

$\Rightarrow$ easiest to motivate w/ $L = 2$ case (e.g. 2 gene deletions)

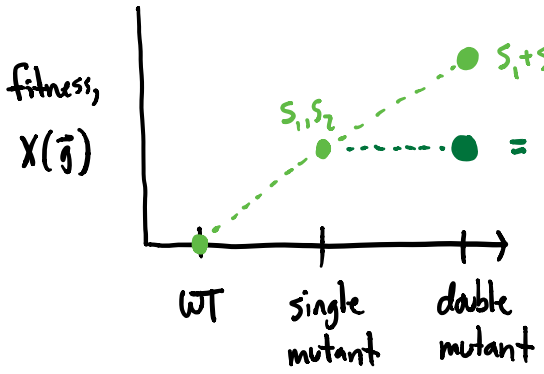
$X(0,0) \equiv 0$ (convention)

$X(1,0) \equiv s_1$
$X(0,1) \equiv s_2$
} could measure, e.g. gene deletion screen (HW2)

$X(1,1) \equiv ? \equiv \underbrace{s_1 + s_2}_{\text{"additive part"}} + \underbrace{\epsilon}_{\text{"epistasis" (how much deviation from additivity)}}$

e.g. "$\epsilon > 0$" $\Rightarrow$ "positive epistasis" $\Rightarrow$ "sign epistasis"
"$\epsilon < 0$" $\Rightarrow$ "negative epistasis" etc. etc.

Often easiest to express w/ picture:

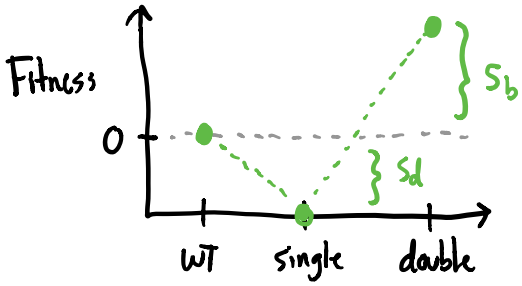


s_1+s_2 = "additive", e.g. orthogonal pathways in a cell

= "diminishing returns", e.g. 2 genes in same pathway

2 stop codons in same gene...

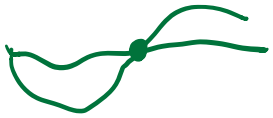
$\Rightarrow$ people often interested in scenarios like:



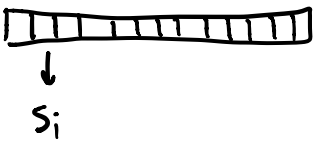
"fitness valley crossing"

e.g. initiation of cancer

contact residues in proteins



$\Rightarrow$ gets even more complicated for $L>2$:



$$X(\vec{g}) \equiv \underbrace{\sum_{\ell=1}^{L} s_\ell g_\ell}_{\text{additive part ("coupon collecting")}} + \underbrace{\epsilon(\vec{g})}_{\text{epistatic part.}}$$

$\Rightarrow$ can write as Taylor expansion around WT:

$$\epsilon(\vec{g}) = \underbrace{\sum_{\ell=1}^{L}\sum_{\ell'=1}^{L} \epsilon_{\ell\ell'} g_\ell g_{\ell'}}_{\text{"pairwise epistasis"}} + \underbrace{\sum_{\ell}\sum_{\ell'}\sum_{\ell''} \epsilon_{\ell\ell'\ell''} g_\ell g_{\ell'} g_{\ell''} + \ldots}_{\text{"higher order epistasis"}}$$

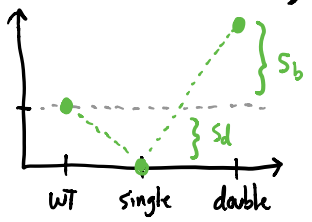
$\Rightarrow$ hard to parameterize in general (active area of research!)

$\Rightarrow$ in practice, people often use:

Additive model ($L \gg 1$)

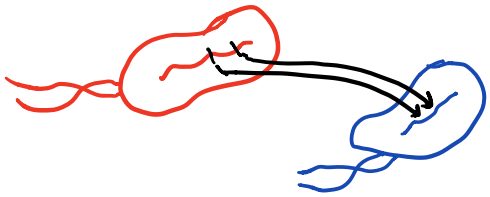
$$X(\vec{g}) \approx \sum_{\ell=1}^{L} s_\ell g_\ell$$

Pictures ($L \sim O(1)$)



⇒ other new bit of biology for $L \geq 2$:

(5) Recombination (exchange of genetic material between different individuals)



Many different mechanisms!

⇒ but many share same basic behavior:

(1) Focal individual (f) is chosen to undergo recombination

⇒ w/ probability ρ per individual per-gen → e.g. mating viruses/phage uptake of DNA cellular DNA, d

(2) Donor individual (d) is chosen to donate portion of genome

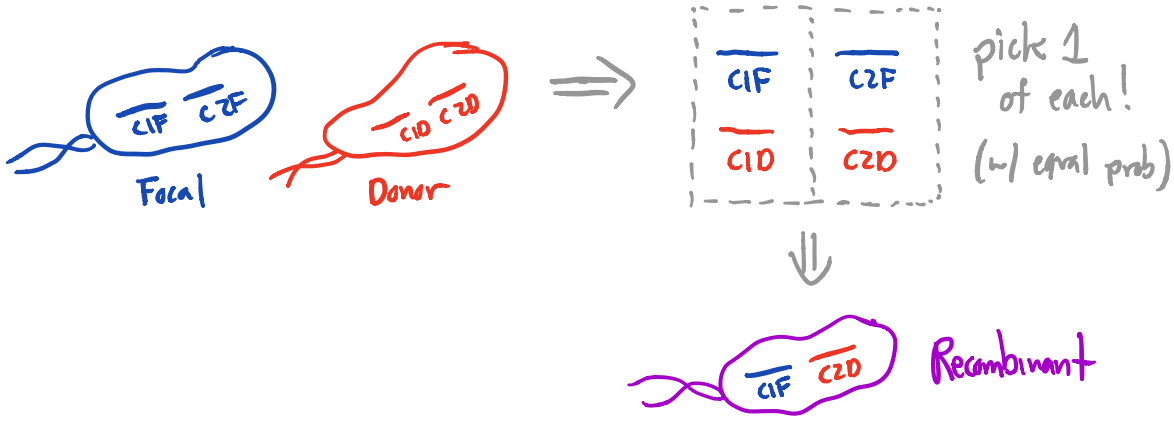
⇒ probability ~ $1/N$ ⇒ $f(\vec{g})$ for any individual of that genotype.

(3) some piece of donor's DNA is integrated into focal genome

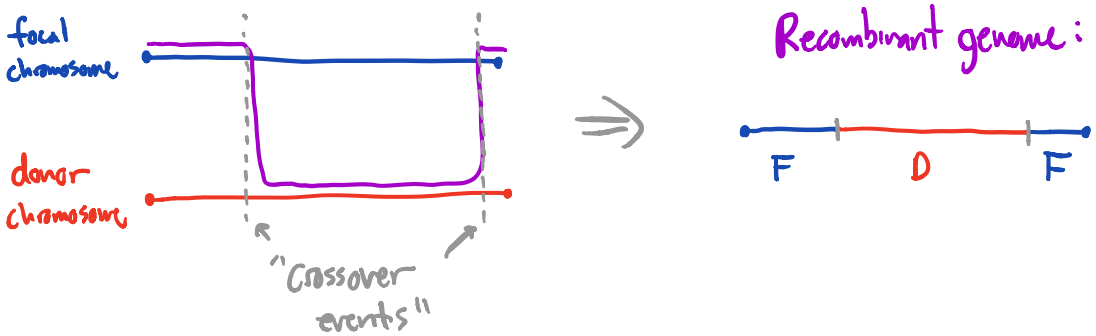
⇒ producing "recombinant"

⇒ different mechanisms enter @ this step:

(a) **Reassortment** (e.g. different chromosomes, e.g. yeast, humans, influenza.)

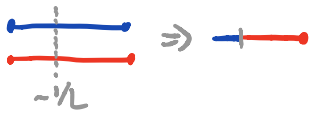


(b) **Crossover Recombination** (e.g. w/in chromosomes in humans)



⇒ often modeled w/ ~1 crossover per recombination event

w/ location chosen uniformly across chromosome



⇒ in practice, "hot spots" & "cold spots" ⇒ "recombination map"

⇒ effective recombination rates vary over many orders of magnitude for different sites in same genome!

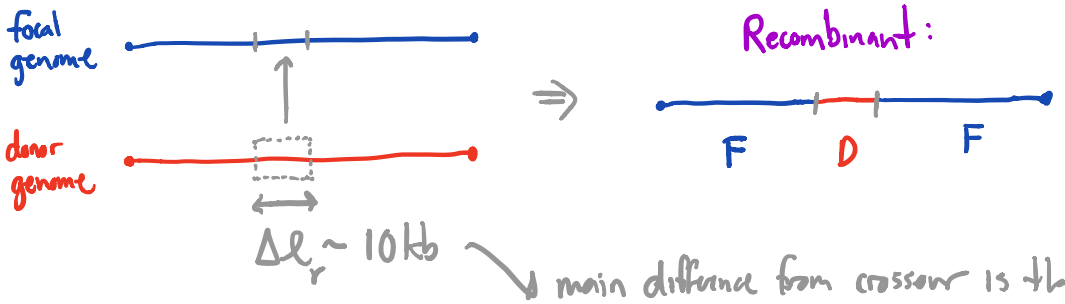
⇒ e.g. in humans ⇒ $L_{chrom} \sim 10^8$ bp

⇒ P(recomb) ~ 100% if 2 ends of same chrom

⇒ P(recomb) ~ 10^{-8} if neighboring base pairs

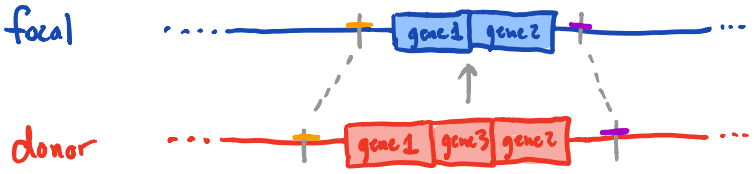
ⓒ "Horizontal gene transfer" / "gene conversion"

⇒ lingo is a little controversial, but basic idea pretty simple:



main difference from crossover is that $\Delta \ell_r \not\propto L$, as opposed to $\Delta \ell_r \sim \mathcal{O}(L)$

⇒ also a mechanism for gaining + losing genes ("accessory genome")

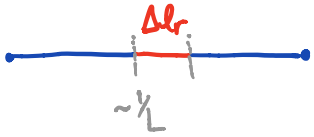


often mediated by homology (=)
similar to PCR...

⇒ active area of research!

⇒ but in this class, will mostly focus on "core genome"

⇒ simplest HGT model:



Δl_r = const, location ~ uniform

So far: individual-based picture...

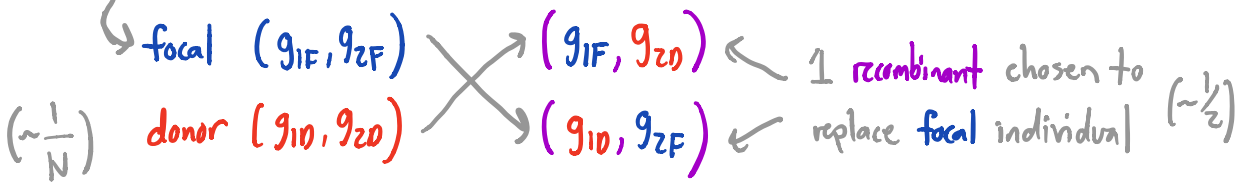
⇒ can we translate to continuum limit?

$$\left(\frac{\delta f(\vec{g})}{\delta t}\right)_{rec} = ???$$

⇒ easiest to start w/ $L=2$ case ⇒ $\vec{g} = (g_1, g_2)$

⇒ all mechanisms have same net effect:

$\Rightarrow$ w/ rate R [function of $\rho, L, \Delta_{lr}, \ldots$ etc.]



$\Rightarrow$ total outflow from recombination: $-Rf(\vec{g})$

$\Rightarrow$ total inflow? $2^2 \times 2^2 = 16$ possible focal/donor combos

case 1 (of 16):

F $(1,1)$ $\rightarrow$ $(1,0)$ $\Rightarrow$ rate $Rf(1,1)f(0,0)\cdot\frac{1}{2}$

D $(0,0)$ $\rightarrow$ $(0,1)$ $\Rightarrow$ $Rf(1,1)f(0,0)\cdot\frac{1}{2}$

case 2 (of 16): F $(0,0)$ $\rightarrow$ $(0,1)$ same!

D $(1,1)$ $\rightarrow$ $(1,0)$

Case 3 (of 16): $(1,1) \to (1,0) \Rightarrow Rf(1,1)f(1,0)\,\frac{1}{2}$

$(1,0) \to (1,1) \Rightarrow Rf(1,1)f(1,0)\,\frac{1}{2}$

$\Rightarrow$ after tabulating all 16 combinations (all 32 recombinants) can add them up to obtain:

$$\left(\frac{\delta f(1,1)}{\delta t}\right)_{rec} = Rf(1,0)f(0,1) - Rf(1,1)f(0,0)$$

$$\left(\frac{\delta f(0,0)}{\delta t}\right)_{rec} = Rf(1,0)f(0,1) - Rf(1,1)f(0,0)$$

same!

$$\left(\frac{\delta f(1,0)}{\delta t}\right)_{rec} = Rf(1,1)f(0,0) - Rf(1,0)f(0,1)$$

$$\left(\frac{\delta f(0,1)}{\delta t}\right)_{rec} = \text{same.}$$

$\Rightarrow$ normalized so that $\sum_{\vec{g}} \delta f(\vec{g})_{rec} = 0$ ✓

⇒ harder to write down explicitly for $L > 2$ ....

but will have general form:

$$\left(\frac{\delta f(\vec{g})}{\delta t}\right)_{rec} = \rho \sum_{\vec{g}_F, \vec{g}_0} T(\vec{g}_F, \vec{g}_0 \to \vec{g}) f(\vec{g}_F) f(\vec{g}_0) - \rho f(\vec{g})$$

incoming recombinants

outgoing recombinants.

"recombination kernel" → "tensor"

nonlinear!

⇒ unlike mutation, can create genotypes far from $\vec{g}$!

Putting everything together, general multilocus model looks like:

$$\frac{df(\vec{g})}{dt} = \underbrace{\left[X(\vec{g}) - \bar{X}(t)\right] f(\vec{g})}_{\text{selection (nonlinear)}} + \underbrace{\sum_{\vec{g}'} M(\vec{g}' \to \vec{g}) f(\vec{g}') - M(\vec{g} \to \vec{g}') f(\vec{g})}_{\text{mutation (linear, "local")}}$$

$$+ \underbrace{\rho \sum_{\vec{g}_F, \vec{g}_D} T(\vec{g}_F, \vec{g}_D \to \vec{g}) f(\vec{g}) - \rho f(\vec{g})}_{\text{recombination (nonlinear, non-local)}}$$

$$+ \underbrace{\sqrt{\frac{f(\vec{g})}{N}}\,\eta(\vec{g}) - f(\vec{g}) \sum_{\vec{g}'} \sqrt{\frac{f(\vec{g}')}{N}}\,\eta(\vec{g}')}_{\text{genetic drift (stochastic)}}$$

Problem: No exact solution for stationary dist'n, p_{fix}, etc.
— even for $L = 2$!

$\Rightarrow$ What do we do instead?!? $\Rightarrow$ asymptotic approx's

Question: Given parameters ("knobs") $L, N, X(\vec{g}), M, \rho, T$

$\Rightarrow$ what are some limits where we might understand understand this SDE?

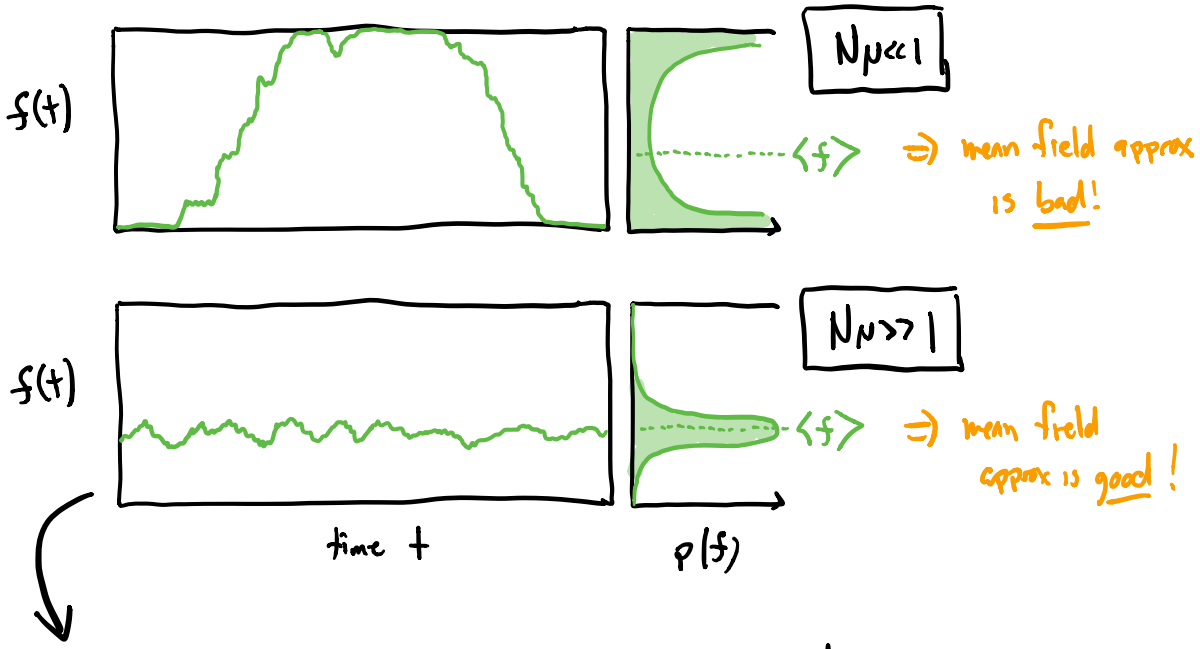
$$\frac{df(\vec{g})}{dt} = \sim(x - \bar{x}) + \sim L\times\mu + \sim\rho + \sim\frac{\pm}{\sqrt{N}}$$

① Obvious answer: $L=1 \Rightarrow$ cheating!*

② in physics, might be primed to take $N \to \infty$ limit ...
("mean field approx") since @ least noise goes away ...

$\Rightarrow$ is this a good approx here?

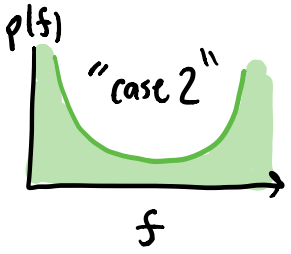
$\Rightarrow$ Recall for $L=1$ case, 2 different regimes when $t \to \infty$:



key feature: large # of individuals in both genotypes @ same time
$\Rightarrow$ so fluctuations are small.

$\Rightarrow$ e.g. for $L=2$, might be ok $\Rightarrow$ but for $L \gg 1 \Rightarrow 2^L \gg N$!
e.g. $L \sim 1000$ bp $\Rightarrow 2^L \sim 10^{300}$!

⇒ large L will always look like
(@ least in some dimensions)



⇒ noise always relevant!

Need to look for other approximations of SDE...

$$\frac{ds(\vec{g})}{dt} = \sim(x-\bar{x}) + \sim L\times\mu + \sim\varrho + \sim\frac{z}{\sqrt{N}}$$

Let's revisit our first idea (L=1)

⇒ even if $L \gg 1$, if behavior "looks like" L=1 case,
⇒ can use what we already know...

(3) Successive mutations regime (i.e. treat mutation as small correction)

⇒ what if mutation rates are low enough that
only 1 or 2 genotypes are present @ a time?